INFORMATION
SYSTEMS IN
MANAGEMENT

Information Systems in Management (2014) Vol. 3 (4) 250–260

ANALYSING AND PROCESSING OF GEOTAGGED SOCIAL MEDIA

PIOTR OLEKSIAK

Department of Applied Informatics, Faculty of Management and Economics, Gdańsk University of Technology

The use of location based data analysing tools is an important part of geomarketing strategies among entrepreneurs. One of the key elements of interest is social media data shared by the users. This data is analysed both for its content and its location information, the results help to identify trends represented in the researched regions.

In order to verify the possibilities of analysing and processing of geotagged social media data, application programming interfaces (APIs) of social networks were examined for their ability to generate reports from the collected data.

The first results of using the system have indicated the possibility of collecting and analysing information generated by Twitter users in real time. Trends and geographical distribution in time can be observed. Further research showed that comparing results and further processing was possible.

Keywords: Geomarketing, Geolocation, Twitter, Social media

1. Introduction

Mobile devices are transforming the way we spend our time, consume information and communicate. Smartphone penetration is growing and becoming dominant with 67% of US mobile subscribers owning a smartphone and monthly time spent by users on mobile devices being over 41 hours in the United Kingdom and 34 hours in the USA as of December 2013. The trend amongst smartphone users indicates that almost 30% of time spent using mobile applications is spent on

social networks such as Facebook or Twitter [4]. This trend is visible also in monthly active users statistics, with Facebook and Twitter having 79% and 78% of monthly active users on mobile devices in the first quarter of 2014 [5].

Social networks share, among others, three publicly articulated features – profiles, friends and comments [1], these turn social media users into influencers enhancing their role in the commercial marketplace [9]. This is the reason social media sites are of interest to business owners and marketing specialists. This is also a reason why social media platforms are interesting to researchers worldwide.

While the prices of data transfer, data storage and smartphones themselves fall the performance of bandwidth enables faster collection and transfer of data to facilitate richer connections [5]. Furthermore additional sensors in mobile devices enable interactions not available to users a couple of years ago. GPS units coupled with network location provide reliable location sources for users, which are used not only in navigation apps but find applications in numerous other mobile software, such as social networking. Content creators use social networks to provide a given photo, tweet or video to the community. This content may sometimes be accompanied by physical location information of the content creator. The content may also be the location of the user, who is sharing his or hers position. Analysing geotagged social media data allows the researcher to define trends in a region, at a specific time or check where phenomena are developing worldwide.

This paper contributes a method of verifying possibilities of analysing and processing geotagged social media data. This method would find application both in business and research institutions enabling spatial analysis of social media data, Twitter in particular.

This paper will proceed as follows: Section 2 introduces social networks incorporating geolocation; Section 3 presents a possible method of gathering geotagged social media data. Section 4 presents results of gathering geotagged social media data, the author's finding and limitations. In Section 5 the author concludes that social media data has research and business application potential and how improvements to the method could be applied.

2. Geotagged social media data

Data uploaded by users of social media platforms such as Facebook, Twitter or Instagram sometimes contains more information than just the post, photo or tweet itself. This data may be accompanied by volunteered geographic information, which specifies the location where the specific data was uploaded. Geotagged content is also available through specialized geolocation applications such as Foursquare or Yelp. Obtaining this data is possible through the use of GPS

equipped smartphones, which send location information together with the content. As presented in Fig. 1 the content creator uploads, tweets, text, videos or photos together with the location information to the content provider. The content sent by the content creator could even be just his or her location. This is sometimes the case on Facebook or Foursquare, when users check-in to locations and share this information.

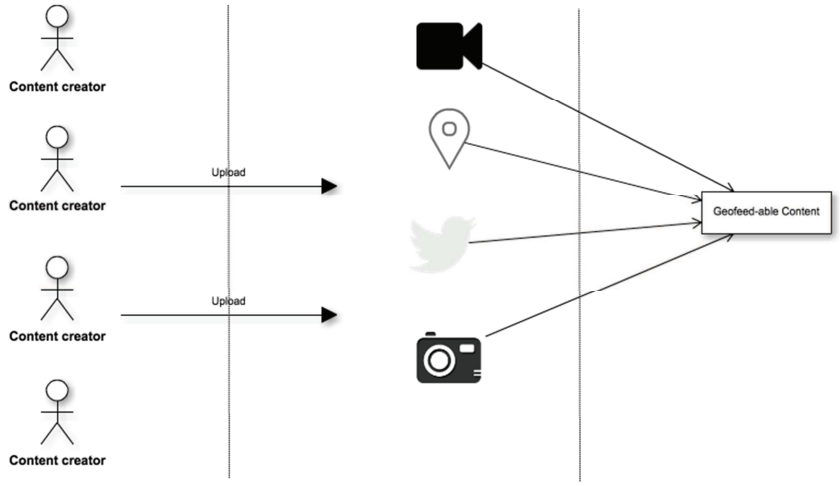

Figure 1. Uploading of geotagged content by content creators to social media networks. *Source*: own preparation on the basis of (Mitchell & Harris, 2013) [6]

Standard geotagged data will consist of typical content for a given social network, however it will be accompanied by a location marker and geolocation information for the feed. In Fig 2. an example of a geotagged tweet is presented with the hashtag #CzasDecyzji, which was used after the polish elections in November 2014. The tweet consists of the name of the content provider, the text itself, the time of publishing and of the location, where the tweet was taken. In this example Toruń, Polska.

The availability of volunteered geographic information is vast. Twitter alone stands for 284 million active users with 77% of accounts outside of the U.S.. These users generate 500 million tweets per day [12]. Not all this data is geotagged, however as Morstatter et al. finds 1 – 3% of all tweets have longitude and latitude bearings [8] giving a total of 5 – 15 million daily geotagged tweets from Twitter alone. Other social networks provide further geotagged data for research or business analysis.

Figure 2. Geotagged tweet. Source: ("WHERE," 2014) [13]

There are possibilities of utilizing geotagged social media data for business and research applications. With the advancement of positioning systems and free access to cartography such as Google Maps, the way in which the society use maps has changed. Web-based maps, in contrast to their static ancestors, are characterised by the democratization of content creating. Everybody can participate in map creating, which allows users to provide cartographic data [14]. Volunteered geographic data, provided through social media networks, is one way of providing this cartographic data and enhancing maps and geographic information systems with trends and location based preferences of users.

Zook and Poorthuis in a study concerning the popularity of beer in different regions of the United States analysed publicly available geotagged Twitter data. With a database collected from June 2012 to May 2013 containing 1 million geotagged tweets a study was conducted concerning the popularity of wine and beer. The popularity of "light" beers and regional beers in the United States was also examined. The data collected from Twitter about the popularity of certain brands in regions coincided with actual sales statistics of these beers in these regions and the activity of Twitter users correlated with the actual behaviour of consumers in the regions [14]. The use of geolocation techniques is accessible to a growing number of people, who can utilize the data for analysis, which was until now available only to GIS experts. Research based on users' activities and location is applicable to many aspects of life. A research conducted on football teams in Great Britain compares the popularity of two rival teams. Manchester United and Manchester City have fans all over the United Kingdom, but also in the city of Manchester.

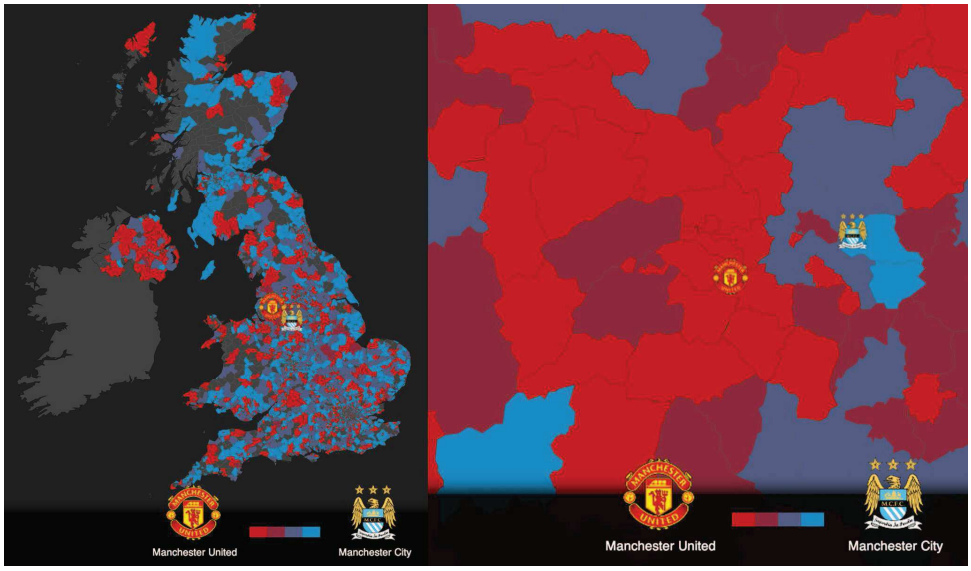

Figure 3. The location of tweets from Manchester United and Manchester City supporters in the U.K. *Source*: Own preparation on the basis of ("floatingsheep: Premier League teams on Twitter (or why Liverpool wins the league and the Queen might support West Ham)," 2013) [2]

Fig. 3 represents the split of the representation of supporters of both teams on Twitter taking into account their location [2]. The blue colour on the figure represents a more densely populated Manchester City supporter region and the red colour represents dominant tweets from Manchester United fans.

3. Gathering geotagged social media data

Both commercial location-based services and location sharing services mostly have been implemented for specific platforms [10]. Most of these services provide APIs, which can be used to retrieve information from their systems, therefore making them content providers for researchers and business analytics wishing to analyse the available data. Twitter data is most commonly used for numerous studies, from finding eyewitness tweets during crises [7] to connecting social and the special networks [11]. Other social networks are also known to be studied for geolocation data, Foursquare and Flickr being an example [3]. All of the content providers mentioned above and others, such as Instagram, provide APIs suitable for geotagged content retrieval. Not all open APIs are however suitable for research purposes as they are restricted both territorially and by the number of

requests a server can make to the given API. Twitter provides two types of APIs, the streaming API and the Firehouse API. Twitter's Streaming API has been used for social media and network analysis to generate understanding of how users behave. The streaming API retrieves 1% of all tweets whilst the Firehouse as much as 43% of all tweets [8]. The downside of the Firehouse is the cost of infrastructure and bandwidth as well as data storage, which for such amounts of data are critical. In Fig 4. a schematic diagram of the geotagged data retrieval system is presented. This type of system is used to retrieve geotagged social media data from any content provider, such as Twitter and present the results to the end user.

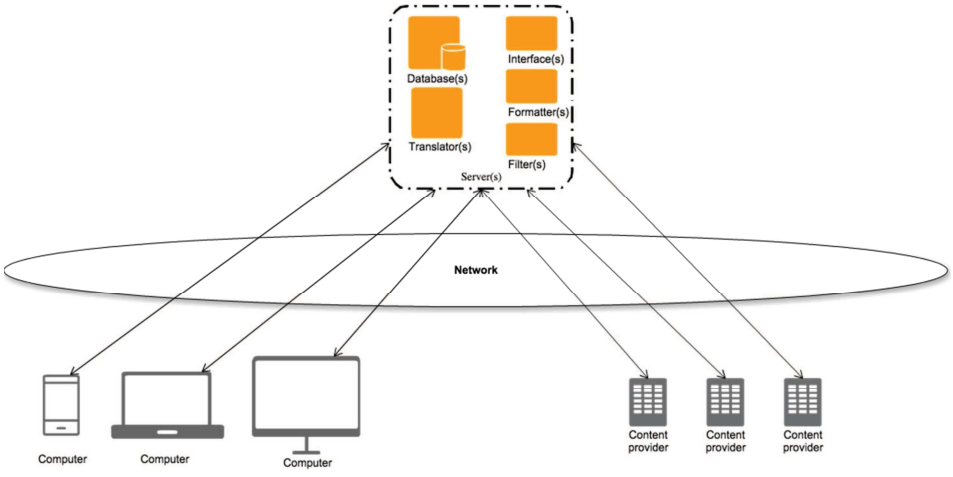

Figure 4. Schematic diagram of the geotagged data retrieval system. *Source*: own preparation on the basis of (Mitchell & Harris, 2013) [6]

The system itself consists of a server or multiple servers with databases, interfaces, formatters, translators and filters. It communicates with the content providers through the network sending requests to the APIs and retrieving desired data. For example a request could ask for all tweets with the hashtag #electionday in a given area. Requests could consist of multiple texts or hashtags as well as multiple areas. The request is typically user defined based on the needs of the analysis. The data is stored in the databases and filtered depending on the needs of the system. A typical filtering process could include filtering for geotagged content, filtering multiple content from single users and determining dominant urban area content ratios. The system can then format and display the results in a suitable form. Tables, clusters on maps and graphs are most common visualization types. Formatting and translating conducted on the server includes aggregating data to a common derivative. For example time and date from all sources should be

displayed in one format. Filtered and transformed data is presented usually via a webpage or mobile application to the end user, using mobile devices or computers. The end user may interact with the data and the interface to amend and display what is necessary. This standard geotagged content retrieval system can be used to retrieve any data from social media platforms providing they share their API with the researchers.

4. Discussion and limitations

The system for aggregating geotagged social media data discussed in section 3 was implemented in the study of verifying the possibilities of analysing and processing such data. The first step of the research consisted of gathering information about three social data system's APIs, which could be used for further research. Instagram, Twitter and Foursquare were selected.

Instagram is a photo sharing social network. Users may accompany each post with the exact location taken from GPS and networks' locations. The position is accurate and is both displayed in the application and transferred throught the API as coordinates. Unfortunately research has proven that the open API only gives access to single requests, in example: locations of a given user. This does not allow searching for geographically specified content and, as a result, Instagram was not used in the research.

Foursquare was the second examined content provider and, as Instagram, Foursquare provided accurate location for the users posts. Here the names of each location were also available. Foursquare API for streaming location based check-ins is closed and access was not granted upon request. Foursquare was also not used in further work.

The streaming API of Twitter proved to be the best solution for creating a required system. The infrastructure was set up and filters were only set to retrieve tweets with set coordinates. Other filters were not put in place, as the purpose of the exercise was to check, which filters could be useful for further research. Having set up the system, tests for retrieving the data were conducted and finally two tests were generated to check the benefits and limitations of research conducted through such a system.

The first conducted test aimed at searching for tweets including hashtags and text concerning Real Madrid football club. Hashtags such as #RealMadrid and #HalaMadrid were defined. Data was collected from Monday, November 3rd to Tuesday, November 25th, with a total of 3159 geottaged tweets collected. A dataset in the user interface consisted of the id of the tweet, latitude, longitude, date and time the tweet was taken, full text of the tweet and a direct to the tweet.

Further data such as text and hashtag searched or content creator name are stored in the database. The data distribution was in this case further shown on Google Maps Fig 5.

Figure 5. Data distribution of geotagged tweets. *Source*: ("WHERE," 2014) [13]

From the distribution several phenomena could be observed. First of all results showed where Real Madrid tweets were present, including which agglomerations and parts of the world. Secondly results showed where the intensity of tweets was greatest. Thirdly you could pinpoint the exact location of a specific tweet. Not surprisingly most tweets were sent from the city of Madrid, however other clusters of Real Madrid fans could be found in other parts of the world.

Another presentation tool in the system is the graph tool, which allows the user to choose a period of time to present a graph for the number of geotagged tweets recorded each day. Fig.6 presents a graph for the analysed search. Here you can observe that tweets are much more frequent during match days. Further studies indicate that tweet peaks occur during breakthrough moments in games such as goals.

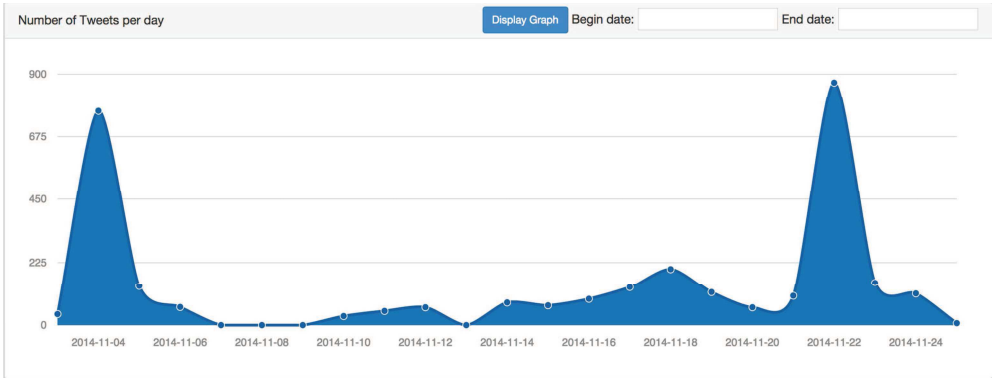

Figure 6. Graph distribution of geotagged tweets. *Source*: ("WHERE," 2014) [13]

The second analysis conducted was aimed to check the polish national Twitter base. The election time was chosen with the hashtag #CzasDecyzji. The results showed 129 tweets, most just after the elections. This is a much smaller group for analysing data however three main conclusions come from this research. Rural use of geotagged social media data is much smaller than its urban equivalent. Heavy users can dominate the discussion and filters should be put in place to block multiple tweets. As many as seven tweets from one user were noted during the study. Twitter use in Poland is scarce compared to other regions and filters would need to be put in place in order to demonstrate a per user style analysis in future research.

The aggregated data proves that geotagged social media photos, tweets or videos could be useful for business analysis and research as anticipated. The analysis is possible and could be used for larger datasets as well as specified feeds. Although not showed in the presented work this type of analysis would also be of use when trying to show trends in a given region. In this scenario an area would be selected and all geotagged data from this region would be selected. This type of analysis would be helpful in detecting trends and sudden anomalies, which could be useful for both business and emergency institutions inspecting live activities in a selected region.

Although geotagged social media data can be applicable to analysing trends, this analysis, as proven, has many limitations. One of the limitations shown is a bias trend towards urban areas. Technology is consumed differently in rural and urban areas and this is also the fact in geolocation social media use. Secondly heavy users tend to dominate discussion on Twitter, which may distort the results of a study, if not filtered. Thirdly to prevent domination, international research would need to take into account the use of a given social network in each country. Furthermore these analysis result are only limited to the users of social networks and do not take into account the whole population. Lastly, which may not be considered as an analytical problem, but could be seen as a privacy issue, the researcher may find the exact location of a tweet author.

5. Conclusions

The created system for gathering geotagged social media data utilised Twitter's streaming API in order to aggregate data provided by the content provider. The assumption that analysis of trends occurring in selected regions could be performed through analysing geotagged social media data proved to be correct.

Data collected during the study of tweets, which were accompanied by location bearings, was analysed. The presentation was conducted in the form of

marker clusters on a map, a tabular summary of all collected data and a graph showing the number of tweets in a selected time frame.

The results confirm that the occurrence of geotagged tweets are spatially dispersed and that real time analysis is possible. This was especially visible in the research of Real Madrid related tweets, where supporters of the team could be grouped into regions with the dominance of Spain's capital. The time of tweets was highest during game days, especially during scored goals. Further analysis of this data could lead to interesting conclusions, however limiting the limitations of the system, such as urban bias or heavy user tweets, should be conducted through filtering.

Geotagged social media data is openly available and provides information, that used to be available only for GIS experts, therefore the created system may be of high value both for business analysis and research of trend analysis.

REFERENCES

[1] Boyd, D. (2007). Why youth (heart) social network sites: The role of networked publics in teenage social life. ... *Series on Digital learning–Youth, Identity, and Digital ...*, *7641*. doi:10.1162/dmal.9780262524834.119

[2] floatingsheep: Premier League teams on Twitter (or why Liverpool wins the league and the Queen might support West Ham). (2013). Retrieved April 21, 2014, from http://www.floatingsheep.org/2013/01/premier-league-teams-on-twitter-or-why.html

[3] Hecht, B., & Stephens, M. (2014). A Tale of Cities : Urban Biases in Volunteered Geographic Information. Retrieved from http://www.users.cs.umn.edu/~bhecht/publications/bhecht_icwsm2014_ruralurban.pdf

[4] How Smartphones are Changing Consumers' Daily Routines Around the Globe. (2014). Retrieved September 01, 2014, from http://www.nielsen.com/us/en/insights/news/2014/how-smartphones-are-changing-consumers-daily-routines-around-the-globe.html

[5] Meeker, M. (2014). *Internet trends 2014-code conference*. *Retrieved May*. Retrieved from http://www.kpcb.com/internet-trends

[6] Mitchell, S., & Harris, P. (2013). System and Method for aggregating and distributing geotegged content. United States.

[7] Morstatter, F., Lubold, N., & Pon-Barry, H. (2014). Finding Eyewitness Tweets During Crises. *arXiv Preprint arXiv: ...*. Retrieved from http://arxiv.org/abs/1403.1773

[8] Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *ICWSM*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewPDFInterstitial/6071/6379

[9] Riegner, C. (2007). Word of Mouth on the Web: The Impact of Web 2.0 on Consumer Purchase Decisions. *Journal of Advertising Research*, *47*(4), 436. doi:10.2501/S0021849907070456

[10] Rost, M., Cramer, H., Belloni, N., & Holmquist, L. (2010). Geolocation in the mobile web browser. *Proceedings of the 12th …*, 423. doi:10.1145/1864431.1864468

[11] Stephens, M., & Poorthuis, A. (2014). Connecting the social and the spatial networks on Twitter. *Computers , Environment and Urban Systems Follow Thy Neighbor*.

[12] Twitter- About the company. (2014). Retrieved November 24, 2014, from https://about.twitter.com/company

[13] WHERE. (2014). Retrieved November 24, 2014, from http://whereproject.com/

[14] Zook, M., & Poorthuis, A. (2014). Offline Brews and Online Views: Exploring the Geography of Beer Tweets. In *The Geography of Beer* (pp. 201–209). Springer Netherlands. doi:10.1007/978-94-007-7787-3_17